Decision Making in Uncertain and Changing Environments*

Karl H. Schlag† Andriy Zapechelnyuk‡

June 18, 2009

Abstract

We consider an agent who has to repeatedly make choices in an uncertain and changing environment, who has full information of the past, who discounts future payoffs, but who has no prior. We provide a learning algorithm that performs almost as well as the best of a given finite number of experts or benchmark strategies and does so at any point in time, provided the agent is sufficiently patient. The key is to find the appropriate degree of forgetting distant past. Standard learning algorithms that treat recent and distant past equally do not have the sequential epsilon optimality property.

Keywords: Adaptive learning, experts, distribution-free, ε-optimality, Hannan regret

JEL classification numbers: C44, D81, D83

* The authors thank Sergiu Hart, Gábor Lugosi and Ander Pérez Orive for valuable comments. Karl Schlag gratefully acknowledges financial support from the Department of Economics and Business of the Universitat Pompeu Fabra, Grant AL 12207, and from the Spanish Ministerio de Educacion y Ciencia, Grant MEC-SEJ2006-09993.

† Department of Economics and Business, Universitat Pompeu Fabra, Ramon Trias Fargas 25−27, 08005 Barcelona, Spain. E-mail: karl.schlag@upf.edu.

‡ *Corresponding author.* University of Bonn, Kyiv School of Economics and Kyiv Economics Institute. E-mail: zapechelnyuk@hcm.uni-bonn.de

1 Introduction

Real-life processes are very complex, and even a mathematician who is skilled in computing optimal strategies may find decision making in a natural environment to be a daunting task. People often cope with such tasks by seeking advice of experts, imitating their peers or business partners. This typically does not solve the problem as the amount of advice one receives seems to increase in the complexity of the environment. The choice is shifted to a different level, to decide whose advice to follow. Given that the environment is constantly changing, the problem is further complicated, as one wants to be flexible enough to switch to a different expert if there is a sign that the current one is not providing the best advice any more. Flexibility has to be sufficient in order to prevent the decision maker from wishing to abandon the strategy in favor of a different one after a particular, possibly unlikely sequence of events. So one needs strategies that are sequentially rational, much in the spirit of focusing on subgame perfection instead of Nash. There exists an extensive literature both in machine learning[1] and economics[2] that provides simple learning algorithms for natural environments. However, we show that these are not sequentially rational. So the question of existence of a simple algorithm remains.

The environment considered in this paper is as follows. A decision maker (for short, Agent) repeatedly makes decisions in an unknown environment (Nature). In every discrete period of time Agent chooses an action and, simultaneously, a state of Nature is realized. Agent's payoff in a given period depends on her action, as well as on the realized state. We assume that all past states are observable by Agent. Agent can thus compute the payoff that would have been realized by each action in each past period, a scenario also referred to as learning under "foregone payoffs" or "full information".[3] Agent has no prior beliefs about Nature's behavior: it may be as simple as a deterministic sequence of states or a stationary stochastic process, or

[1] Littlestone and Warmuth (1994); Cesa-Bianchi et al. (1996); Vovk (1998); Auer and Long (1999); Foster and Vohra (1999); Freund and Schapire (1999); Cesa-Bianchi and Lugosi (2003, 2006); Greenwald and Jafari (2003); Cesa-Bianchi et al. (2007); Gordon et al. (2008).

[2] Hannan (1957); Foster and Vohra (1993, 1997, 1998); Fudenberg and Levine (1995, 1999); Hart and Mas-Colell (2000, 2001a); Lehrer (2003); Hart (2005).

[3] In Section 8 we show how to extend our analysis to the multi-armed bandit setting where only own payoffs are observable.

as complicated as strategic decisions of a hostile player who seeks to inflict Agent maximum harm. So Agent is trying to learn in a distribution-free environment.

We do not aspire to find the first best strategy for Agent. In fact, this is an impossible task if one does not add priors, which is equivalent to adding structure on the environment. Since Nature's complexity is unbounded, even a very patient Agent cannot hope to "learn" Nature's behavior. Instead, we wish to find a strategy so that Agent performs as well as those surrounding her that are facing the same environment. These can be *experts* that are making recommendations to Agent, other agents that are also making choices, or simply strategies that Agent considers as benchmarks. In what follows we summarize these three entities in the term *expert* and assume that these experts are given and finite in number. It is important that we allow Agent to observe past states so that the past performance of each of these experts can be evaluated.[4] The objective of Agent is to perform similarly to the best of the experts without prior knowledge which expert is actually the best.[5] That is, she wishes to guarantee that the expected sum of the discounted future payoffs is close to or above that of each expert. Moreover, Agent aims to achieve this objective not only in the first period, but at any point in time. So, we search for a strategy that is *dynamically consistent*. This prevents Agent from choosing some strategy in period 1 and then changing her mind at some later time after a particular sequence of events (thus precluding the problem of choosing some strategy when knowing in advance that it will not be carried out). Moreover, Agent will also prefer not to change her strategy after she has made a mistake. This is just the standard condition of *sequential rationality* (or subgame perfection) that demands optimality of a strategy after every history – including those that have zero probability.

We find that a strategy need not be very complex to achieve this objective. We design a simple learning algorithm for Agent that guarantees the expected sum of the discounted future payoffs to be ε-close to that of the best of the experts, consistently in all periods of time, regardless of Nature's behavior. Furthermore, we show that Agent can approach the performance of the best expert arbitrarily closely, provided she is sufficiently patient. The algorithm is described as follows. In every period,

[4]Alternatively, one can assume that Agent does not observe past states but instead observes own past payoff as well as those of all experts (see also Section 7).

[5]In fact, different experts may be best in different periods.

Agent assesses the past performance of each expert (a weighted sum of the payoffs that Agent would have gotten if she always followed that expert's advice in the past). Then Agent follows an expert's advice with probability proportional to how much better that expert performed in the past relative to Agent herself, similarly to Hart and Mas-Colell's *regret matching* strategy (Hart and Mas-Colell, 2000, 2001a).[6]

The key to our strategy designed for Agent is the way in which the past performance of experts is assessed. Unlike Hart and Mas-Colell (2000), where all past periods count equally, here Agent puts higher weights on more recent events, regarding more distant events and associated foregone payoffs as less relevant. Though this way of treating the past has been well documented in the psychology literature as the *recency effect* (see Ray and Wang 2001 and the references within) and has been used in a few papers (Roth and Erev, 1995; Erev and Roth, 1998), here this has a strategic reason. The ability to gradually forget the past helps Agent to adapt to changing environments. In contrast, incorporating all past events equally makes the strategy too inflexible, and, indeed, we show that the regret matching strategy of Hart and Mas-Colell (2000) does not satisfy the sequential rationality property.

It is important to note that Agent herself cannot compute expected future payoffs neither for her strategy nor for the experts, since she does not know Nature's behavior; computation is possible only from an observer's point of view. Yet, with our algorithm Agent can make a comparative statement about her expected future payoffs relative to the experts'. We provide a bound on how much Agent's expected payoffs can differ from that of the best expert and show that Agent can perform arbitrarily close to or better than the best of the experts provided she is sufficiently patient. We also extend this result to the setting where we allow for errors in observing outcomes.

This paper is different from the existing literature in three aspects. The first aspect relates to the richness of our setting. The set of Agent's actions, as well as the set of states of Nature, need not be finite, as opposed to those in finite-game models such as Fudenberg and Levine (1995, 1999); Hart and Mas-Colell (2000, 2001a). Agent's utility function need not be linear or convex, and the experts need not play deterministic strategies, as it is assumed throughout the machine learning

[6] Alternatively, Agent chooses a convex combination of the experts' recommendations with weights proportional to the correspondent differences in performance, if Agent's action space is convex and her utility function is concave.

literature.

The second difference from the literature concerns the objective that we specify for Agent. Future payoffs are discounted in line with classic decision theory. In each period these cumulated payoffs are compared to those of the experts. In contrast, the existing literature uses time-averaging and evaluates payoffs from the perspective of the first period only (see Cesa-Bianchi and Lugosi, 2006, and references within). Furthermore, we compare *expected* payoffs of strategies used by Agent and experts while the existing literature compares realized payoffs and establishes almost sure bounds. For better comparison to this literature we formulate our results in terms of probabilistic bounds in Appendix B.

In fact, Agent's discount factor plays a novel role in this setting. A less patient Agent has higher goals as she aspires to achieve higher period-by-period payoffs. The reason is that Agent wishes to do as well as the best expert. Payoffs accumulated from following the best expert in each short run will be higher than that from following the single best expert in the long run. But, of course, a less patient Agent has greater difficulties in learning, as she needs to learn which expert is best in each short run. Depending on which effect is greater, from the viewpoint of an outside observer, a more patient agent may or may not perform on average better than a less patient one.

The third difference of our paper from the literature is that we achieve our objective by conditioning future choices on a weighted assessment of past payoffs, putting larger weights on more recent periods. In contrast, practically all strategies found in the literature condition future play on time-averages of the past performance. As we show in this paper, they thus lack the property of dynamic consistency and hence cannot guarantee Agent's sum of discounted future payoffs to be close to that of the best expert in all periods. The problem of time averaging of the past is that it eventually leads to an inability to react to changes in the environment. As time passes, a decision maker adds smaller and smaller weights on new observations and thus requires increasingly large body of evidence to change her opinion once it is settled. So, a decision maker who treats past events equally is likely to end up in a situation where in response to a changing environment she would prefer to "forget" all the past and start afresh, with an empty history, rather than to continue using the original

strategy.

There are a few papers that previously considered discounting of past payoffs. Roth and Erev (1995) and Erev and Roth (1998) use reinforcement learning models with a small degree of "gradual forgetting" to explain experimental data on some simple games, such as the ultimatum bargaining game. Cesa-Bianchi and Lugosi (2006) consider maximizing discounted past payoffs as Agent's objective (while we use this assessment of previous performance only to determine Agent's future play). Marden et al. (2007) study a special class of finite games that are acyclic in better replies and show that if all players play strategies based on discounted past payoffs with inertia, their play converges to a Nash equilibrium.

The paper is organized as follows. We begin with a motivational example (Section 2). The model is described in Section 3. In Section 4 we introduce strategies based on past payoffs and state our main result. Section 5 discusses the role of adaptation in Agent's behavior and highlights what happens when there is too little adaptation (as in models that condition on time-average payoffs) or too much adaptation. In Section 6 we discuss the role of Agent's discount factor. Section 7 expands the main result to noisy environments. Section 8 concludes. All proofs omitted in the text are deferred to Appendix A. In Appendix B we derive probabilistic bounds on realized discounted future payoffs.

2 Motivational Example

Let us start with a brief motivational example. Consider an investor who trades on a stock exchange and makes a portfolio rebalancing decision once a week. There are various possibilities how the investor can make decisions. She may follow the lead of some respectable company and hold the same portfolio; she may choose to use one of a variety of analytical tools for evaluation of the future dynamics of the financial market, applying it to information obtained from diverse sources. Whose lead to follow? Which analytical tool to use? Which source of information to trust? These are the questions that the investor needs to answer.

In our terminology, any basis for decision making (a company whose lead is followed, or an analytical tool in combination with an information source) is called an

expert who provides advice. The task of the investor is to choose which expert to follow in every decision that she makes. Unfortunately, there does not exist (and cannot exist in principle) a universally good expert. Following advice of a particular expert can bring benefit or loss, depending on future states of Nature. Some experts provide the best advice when the economy is steadily growing; others when it is declining; and others when there is a large degree of uncertainty and fluctuations on the stock market.

We assume that the investor has no prior information or beliefs about future states of Nature and about quality of advice of various experts. Yet, we design a strategy for the investor, based on available experts' advice, that yields the expected annual return nearly as high as the best portfolio among those recommended by the experts, steadily over time, provided that the investor is sufficiently patient.

We illustrate our result by the following stylized example. Suppose that the investor has a certain cash fund and three instruments at her disposal. She can write a certain number of binary call options that the S&P 500 ends the week with a growth, binary put options that the S&P 500 ends the week with a decline, or she can keep cash in bank. Assume that each option costs $50,000$ and yields $100,000$ if the event occurs (thus yielding 100% of conditional return), and otherwise expires worthless (a conditional loss of 100%). The bank yields a safe annual return of 5.2% (or 0.1% per week). Short-selling of the instruments is not allowed.[7]

Denote by $x_t(j)$ the fraction of instrument j in the investor's portfolio in period t, where j indicate one of the three instruments, call option, put option, or cash. In every period t the investor receives the return (net of the cost of the portfolio) of $u_t = x_t(call) - x_t(put) + 0.001 \cdot x_t(cash)$ in the event of growth and $u_t = -x_t(call) + x_t(put) + 0.001 \cdot x_t(cash)$ in the event of decline. The present-value payoff of the investor evaluated at some period t_0 is the discounted sum of all future payoffs,

$$U_{t_0} = \sum_{t=t_0}^{\infty} \delta^{t-t_0} u_t,$$

where δ is the investor's discount factor.

[7]Usually, a binary call (put) option would be conditioned on the event that the S&P 500 grows (declines) by x points, $x > 0$. For simplicity we choose $x = 0$ and forbid short sales to prevent arbitrage. One can easily construct a slightly more complex example with $x > 0$ and then also allow for short sales.

Consider the following strategy of the investor. For every period t denote by u_t^j the return in period t of the portfolio that consists only of instrument j, $j \in \{call, put, cash\}$. Next, denote by $C_{\alpha,t}(j)$ the weighted average value of holding the portfolio consisting of instrument j up to period t,

$$C_{\alpha,t}(j) = (1-\alpha)\sum_{i=1}^{t} \alpha^{t-i} u_t^j.$$

Similarly, let

$$C_{\alpha,t}(0) = (1-\alpha)\sum_{i=1}^{t} \alpha^{t-i} u_t$$

be the weighted average of past payoffs of the investor. Thus $C_{\alpha,t}(j)$ is a measure of the value of holding the portfolio consisting of instrument j in all previous periods, putting highest weight on the most recent periods. Similarly, $C_{\alpha,t}(0)$ is a measure of how well the investor has performed. The excess weighting of recent past will be instrumental to ensure good performance of the strategy when the environment is changing.

The strategy prescribes to hold the portfolio with fraction of instrument j proportional to $[C_{\alpha,t}(j) - C_{\alpha,t}(0)]_+ = \max\{C_{\alpha,t}(j) - C_{\alpha,t}(0), 0\}$, that is,

$$x_{t+1}(j) = \frac{[C_{\alpha,t}(j) - C_{\alpha,t}(0)]_+}{\sum_{j' \in \{call,put,cash\}} [C_{\alpha,t}(j') - C_{\alpha,t}(0)]_+},$$

whenever $C_{\alpha,t}(j) \geq C_{\alpha,t}(0)$ for some j, and otherwise chooses an arbitrary portfolio (for instance, keep the one from the previous period). Thus, only recommendations of experts whose performance is evaluated superior to own will be followed, the probability of following the recommendation of any such expert being proportional to how much better he performed.

We show that a sufficiently patient investor (δ close enough to 1) can guarantee an expected discounted future payoff that is arbitrarily close to the best that can be obtained by any portfolio that remains constant over time. This is true from the perspective of any period t, evaluating future payoffs with discount factor δ, no matter what states of Nature will be realized in future. The value $1-\alpha$ can be considered as the *rate of adaptation* of the investor's portfolio, and it has to be fine-tuned to guarantee the best result. If α is too close to 1, then the rate of adaptation is very slow. For example, in the case when a long series of growth is followed by a long

series of decline, it will take the investor a substantial period of time to adapt and cause her to hold a big share of call options in the portfolio for a long time. If α is too small, then the investor reacts to every fluctuation of the events, and her portfolio will be too volatile and susceptible to small fluctuations. As we show later, the right balance dictates to choose $1-\alpha$ to be of the order of $\sqrt{1-\delta}$.

To be more specific, suppose that it turns out that the annual rate of return on the call option is equal to 20%, resulting from the S&P 500 exhibiting a weekly growth $x\%$ more often than a decline. Then the above strategy guarantees the investor the expected annual rate of return $20\% - \varepsilon(\delta)$, where $\varepsilon(\delta)$ converges to zero as the level of the investor's patience, δ, approaches 1. If instead the annual rate of the put option is 20%, then this strategy will yield the same expected annual rate of return, $20\% - \varepsilon(\delta)$. In fact, given such a limited set of instruments, the worst case for the investor is a constant fluctuation of the S&P 500 around zero with no long-run tendency of growth or decline, where the best portfolio is to hold 100% of cash in a bank. In this case the above strategy guarantees the investor the annual rate of return of $5\% - \varepsilon(\delta)$. Thus, this strategy is almost as safe as keeping cash in a bank, yet it allows the investor to obtain much more whenever there exists a portfolio that yields a higher return.

3 Preliminaries

A decision maker (for short, Agent) repeatedly faces an uncertain environment (referred to as Nature). In every discrete period of time $t = 1, 2, \ldots$ Agent chooses an action a_t from a set A of available actions, and, simultaneously, a state of Nature, $\omega_t \in \Omega$, is realized. There are also N *experts* (or benchmark strategies) who, before each period, make recommendations to Agent about what action to choose; expert j recommends an action a_t^j from A in period t. Let u be Agent's payoff function, so $u(a, \omega) \in \mathbb{R}$ is Agent's payoff when choosing action a in state ω. We assume that A and Ω are compact measurable sets (finite or infinite), and $u : A \times \Omega \to \mathbb{R}$ is measurable and bounded. In every period Agent may condition her choice on the recommendations of the experts made for that period as well as on everything that happened in previous periods. There is perfect information about everything that occurred in the past. Specifically, Agent can observe for each past period the ac-

tions chosen by each of the experts as well the state of Nature that occurred. In particular, Agent can derive for each previous period t and each expert j the utility she would have received if she had followed the recommendation of expert j in that period. Denote by $a^e = \left(a^1, ..., a^N\right) \in A^N$ a profile of actions recommended by the N experts, by $h := (a_t, a_t^e, \omega_t)_{t=1}^{\infty}$ a sequence (or path) of actions, recommendations and states, and by $h_t := ((a_1, a_1^e, \omega_1), \ldots, (a_t, a_t^e, \omega_t))$ the history of play up to t. Let $\mathcal{H}$ be the set of all finite histories including the empty history. A strategy of Agent is a map[8] $p : \mathcal{H} \times A^N \to \Delta(A)$ that associates with every history h_{t-1} and every profile of recommendations a^e a randomized action in A to be played in period t. For short, we write $p_t = p(h_{t-1}, a^e)$ for the randomized action chosen by Agent in period t. Similarly, each expert j is endowed with a strategy $p^j : \mathcal{H} \to \Delta(A)$ where $p_t^j = p^j(h_{t-1})$ is the randomized action belonging to ΔA that is recommended in period t by expert j after h_{t-1} has occurred. The state of Nature realized in period t may also depend on what happened previously, formally it is described by a map $q : \mathcal{H} \to \Delta(\Omega)$ where $q_t = q(h_{t-1})$ denotes the randomized state of Nature that occurs in period t conditional on the previous history h_{t-1}. We assume that the utility of Agent is bounded. In fact, all we need is that the set of possible utilities that can be generated by following some expert after some history is bounded. To simplify further exposition, we can transform Agent's utility function affinely so that whenever Agent follows any expert's recommendation, her utility is contained in the interval $[0, I]$ for some $I > 0$.[9]

It is as if Agent faces an opponent, called Nature, that chooses a state based on the strategy q which is unknown to Agent. Agent could be facing a deterministic sequence of states or a stochastic process independent of Agent's actions. Equally, the sequence of future states may depend on past actions of the Agent and of the experts. For instance, it could be that Nature has its own objectives and is engaged in a repeated game with Agent. In particular, we include the case in which Nature knows the strategy p of Agent and is adversarial in the sense that it aims to inflict maximal "harm" on Agent.

The experts have various interpretations. Note that Agent need not know strategy

[8] $\Delta(B)$ denotes the set of probability distributions over a finite set B.

[9] Let $\underline{u} = \inf\{u(p^j(h), \omega) : h \in \mathcal{H}, \omega \in \Omega\}$ and let $I = \sup\{u(p^j(h), \omega) : h \in \mathcal{H}, \omega \in \Omega\} - \underline{u}$. Then replace in the original utility function $u(a, \omega)$ by $(u(a, \omega) - \underline{u})$.

p^j of an expert j. She knows only realizations of j's recommended actions (in the current period as well as in all past periods). Thus, in our setting experts may know more about the environment than Agent does. Some experts may even know Nature's strategy q, though, of course, it does not mean that they will reveal the best actions to Agent. One interesting interpretation is that experts are *forecasters*. An expert makes a forecast of a next-period state of Nature (it could be a point forecast, a confidence interval, a distribution, etc.). Then Agent's problem is to decide which expert to follow, or possibly how to aggregate the forecasts of the different experts. On the other hand, in some applications it is plausible to assume that the strategies p^j of the experts are known by Agent. Such a setting emerges when there are no explicit experts but instead each p^j describes an algorithm, a *benchmark* strategy, that Agent wants to compare her own performance to. This approach is popular in the computer science literature (see Cesa-Bianchi and Lugosi, 2006, and references within). When the set of actions is finite, then it is common in the literature (e.g., Hannan, 1957; Fudenberg and Levine, 1995; Hart and Mas-Colell, 2001a) to consider as benchmarks the set of constant strategies $\{p^a, a \in A\}$ as experts where p^a specifies to play $a \in A$ in every period, irrespective of the history of play.

In this paper we assume that the set of experts or benchmarks is given. How the experts are selected is not considered here (see some comments in Section 8 below).

We would like to note that everything goes through if the set of feasible actions and states are time dependent, $a_t, a_t^j \in A_t$ and $\omega_t \in \Omega_t$ where A_t and Ω_t are endowed with the same properties as A and Ω defined above. Similarly, everything holds if, as in a more classic decision making setting, outcomes are observable while states are not. In this case X is a set of outcomes, $u : X \to \mathbb{R}$ is bounded and $q : A \times \Omega \to \Delta(X)$ is the underlying process that generates outcomes given actions chosen and states realized.

Agent's payoffs accumulated in different periods are combined as in classical decision making by means of discounting. Agent discounts future payoffs with a discount factor $\delta \in (0, 1)$. For given strategies p and q, Agent's *expected utility* at time t_0 is

denoted by $U_{t_0,\delta}(p,q|h_{t_0-1})$ and defined by[10]

$$U_{t_0,\delta}(p,q|h_{t_0-1}) = E\left[(1-\delta)\sum\nolimits_{t=t_0}^{\infty}\delta^{t-t_0}u(a_t,\omega_t)\Big|\, h_{t_0-1}\right]. \tag{1}$$

Note that these expectations only refer to the randomness inherent in p and q. Agent herself does not know q, and hence cannot compute these expectations. We assume that Agent has no prior beliefs about Nature's behavior q (a distribution-free environment). We will be measuring how well Agent's strategies perform in this unknown, possibly, hostile environment. Instead of assigning a prior on Nature's behavior and finding a Bayesian-optimal strategy, or applying some standard non-Bayesian approach, such as the *maximin* objective of finding the best strategy against the worst-case scenario, we consider a very simplistic objective. The objective of Agent is to perform nearly as well as the best expert, regardless of what Nature does and without knowing in advance which expert is actually the best. Moreover, we assume that this objective is maintained after any history. To put it formally, we say that *strategy p is sequentially ε-as good as strategy p'* if for every strategy q of Nature, every period t_0 and every history h_{t_0-1},

$$U_{t_0,\delta}(p,q|h_{t_0-1}) \geq U_{t_0,\delta}(p',q|h_{t_0-1}) - \varepsilon.$$

A strategy p is *sequentially ε-optimal w.r.t. the given experts* if it is sequentially ε-as good as every p^j, $j \in J = \{1,2,\ldots,N\}$.[11] This is the analogue of the concept of contemporaneous perfect ε-equilibrium introduced by Mailath et al. (2005) in the context of repeated games (see also Radner, 1980). Finally, we say that a strategy p is *sequentially ε-optimal* if it is sequentially ε-optimal w.r.t. *any* set of experts.

The requirement that the expected performance evaluated in period t_0 be ε-as good as that of every expert irrespective of the previous history h_{t_0-1} is of particular importance in this paper. On the one hand, this is a *dynamic consistency* constraint on Agent's objective: if Agent decides to choose a strategy p in period t_0, she should

[10] Strategies p and q, together with an initial history h_{t_0-1}, define a stochastic process that determines a probability measure over histories in $\mathcal{H}$; the expectation is taken with respect to that measure. Note that formally the stochastic process depends also on the strategies of the experts, but we omit them in the notations as we assume these strategies are given as a part of the problem description.

[11] An expert's strategy can be treated as the same mathematical object as Agent's strategy, with the property that it does not depend on experts' recommendations.

not change her mind in any period $t > t_0$. A strategy that does not satisfy this constraint would require Agent's commitment at period t_0 to an infinite sequence of future decisions. On the other hand, this is a condition of *sequential rationality* (or subgame perfection) that ensures optimal behavior of Agent even after zero-probability histories achieved by "mistakes" in past decisions of Agent or Nature. In particular, we do not restrict Agent to start with the empty history, the problem is well defined for every initial history, regardless of the way it has been reached.

4 Conditioning on the Past

In this paper we regard Agent as an unsophisticated, non-Bayesian decision maker who uses her past "experience" in a simple way. More specifically, we will consider strategies where decisions of Agent depend in a simple way on own past performance, as well as on that of the experts. Loosely speaking, Agent will choose to follow advice of those experts who performed better than she did. An important part of this paper will deal with how to appropriately measure past performance. Note that this should not be confused with the fact that future payoffs are evaluated using discount factor δ.

The standard in the literature (see Cesa-Bianchi and Lugosi, 2006, and references within) is to condition next choice in period $t+1$ on average past performance (i.e. the arithmetic mean) of self and of each of the experts, averaging over periods from 1 to t. We say that performance is measured using *past average payoffs* if performance up to time t given history h_t is evaluated by its average in periods from 1 to t. Agent's own performance is denoted by $C_{1,t}(0)$ and given by

$$C_{1,t}(0) = \frac{1}{t}\sum\nolimits_{i=1}^{t} u(a_t, \omega_t)$$

performance of expert $j \in J = \{1, \ldots, N\}$ is denoted by $C_{1,t}(j)$ and given by

$$C_{1,t}(j) = \frac{1}{t}\sum\nolimits_{i=1}^{t} u(a_t^j, \omega_t).$$

In this paper we focus on the setting where past performance is measured with "decay", assigning a higher weight to more recent experiences, referred as *discounted past payoffs*. Specifically, for $\alpha \in (0,1)$ and every $j \in J$ define the *past α-discounted*

payoff at period $t = 1, 2, \ldots$ recursively by setting $C_{\alpha,0}(j) = 0$, and for every $t \geq 1$

$$C_{\alpha,t}(j) = \alpha C_{\alpha,t-1}(j) + (1-\alpha)u(a_t^j, \omega_t). \tag{2}$$

To put it differently, $C_{\alpha,t}(j)$ is defined as

$$C_{\alpha,t}(j) = (1-\alpha)\sum_{i=1}^{t} \alpha^{t-i} u(a_i^j, \omega_t). \tag{3}$$

Analogously, the past α-discounted payoff $C_{\alpha,t}(0)$ of Agent is defined.

One may choose to interpret discounting of past payoffs as a decay of past information, an active underweighing of older outcomes as these are perceived as less relevant than recent events. The discounted past payoff, $C_{\alpha,t}(j)$, is an aggregate of the past information, and according to the recursive formula (2), every new piece of information receives the weight of $1-\alpha$ in this aggregate, thus the term $1-\alpha$ can be viewed as Agent's *rate of adaptation* to new conditions. Indeed, large $1-\alpha$ means that Agent places a considerable weight on new information and adjusts the aggregate values fast; $1-\alpha$ close to zero means that Agent places a little weight on new information, and the aggregate values change slowly. In this sense, the evaluation according to past average payoffs can be considered as declining rate of adaptation, the rate of adaptation in period t being equal to $1/t$.

It is worth noting that strategies based on discounted past payoffs are not computationally demanding. Agent need not remember all the past information, she only needs to know the current values of the discounted past payoffs and to update them by the recursive formula (2) in every period.

Consider a strategy p such that for every period t Agent's next-period behavior depends only on her evaluation of the past performance of the N experts as well as on her own past performance. That is, given a vector $x_t \in \mathbb{R}^{N+1}$ consisting of performance measure $x_t(0)$ of Agent and $x_t(j)$ of expert j, $j = 1, .., N$, the next period mixed action of Agent is a function of x_t only: $p_{t+1} = \sigma(x_t)$. Such a strategy p is called a *better-reply strategy* if for every period t, whenever $x_t(j) \geq x_t(0)$ for some $j \in J$,

$$x_t(j) < x_t(0) \quad \Rightarrow \quad p_{t+1}(j) = 0, \quad j \in J. \tag{4}$$

The better-reply property is a natural condition that stipulates to never follow the advice of those experts whose performance is inferior to Agent's own performance.

The related literature in this area has chosen to explain everything in terms of *regret* (see Appendix A for formal definitions). For each expert one computes the regret of not following this expert in a given period as the difference between the payoff of that expert and own payoff. The choice among experts is governed by the average regret of not following recommendations of these experts. The better-reply condition on Agent's strategy means to never follow the advice of an expert that Agent has negative regret for not following his advice in the past. While the interpretations are different, mathematically the two approaches are identical. We provide a few examples that come from this literature.

Example 1 The better reply strategy $p_{t+1} = \sigma(x_t)$ is the *regret matching* strategy (Hart and Mas-Colell, 2000) if the recommendation of expert j is followed with probability proportional to how much better expert j performed than Agent in the past, formally, if $\sigma(x)$ is defined for every $j \in J$ by

$$\sigma_j(x) = \frac{[x(j) - x(0)]_+}{\sum_{k \in J} [x(k) - x(0)]_+} \tag{5}$$

whenever $x(j') \geq x(0)$ for some $j' \in J$, where $[z]_+ = \max\{0, z\}$.[12]

Example 2 More generally, let P be the $l_{\mathbf{p}}$-norm, $P(x) = \left(\sum_{j \in J} x_j^{\mathbf{p}}\right)^{1/\mathbf{p}}$. Then $\sigma(x)$ is called the $l_{\mathbf{p}}$-*norm* strategy (Hart and Mas-Colell, 2001a; Cesa-Bianchi and Lugosi, 2003) if it is defined for every $j \in J$ by

$$\sigma_j(x) = \frac{\nabla_j P([x(j) - x(0)]_+)}{\sum_{k \in J} \nabla P_k([x(k) - x(0)]_+)} = \frac{[x(j) - x(0)]_+^{\mathbf{p}-1}}{\sum_{k \in J} [x(k) - x(0)]_+^{\mathbf{p}-1}}$$

whenever $x(j') \geq x(0)$ for some $j' \in J$. In particular, the l_2-norm strategy is equal to the regret matching strategy. The l_∞-norm strategy assigns probability 1 on experts with the highest performance. It is equivalent to the *fictitious play* (Brown, 1951) if performance is measured using past average payoffs. For large $\mathbf{p}$, the $l_{\mathbf{p}}$-norm strategies based on past average payoffs approximate fictitious play and are called *smooth fictitious play.*[13]

[12] This strategy should not be confused with the regret matching strategy applied to *conditional regrets* that was also introduced by Hart and Mas-Colell (2000).

[13] Fudenberg and Levine's (1995) original definition of smooth fictitious play is different and does not satisfy the better-reply condition (4).

We can now state our main result. For given $\alpha \in (0,1)$ the *regret matching strategy based on past α-discounted payoffs*, denoted by p_α, is the strategy defined at each time t by applying the regret matching rule (5) to the vector of performance assessments given by $C_{\alpha,t}$.

Theorem 1 *For every $\varepsilon > 0$ there exists $\delta_0 \in (0,1)$ such that the following holds. For every $\delta \geq \delta_0$ there exists $\alpha \in (0,1)$ such that p_α is sequentially ε-optimal.*

This result follows directly from Propositions 1 and 2 below. Theorem 1 states that a sufficiently patient Agent can guarantee the expected utility to be arbitrarily close to that achieved by the best of the experts consistently in all periods. This is true without any knowledge about Nature's behavior and without any possibility of assessing ex-ante which expert's strategy is actually the best as measured by discounted future payoffs.

It is important to note that we provide a uniform bound on the difference between discounted future payoffs of Agent and the best expert. This bound is independent of time and history of past play. In contrast, the existing literature (e.g., Hart and Mas-Colell, 2001a; Cesa-Bianchi and Lugosi, 2003) offer strategies based on time-average past payoffs that guarantee Agent's (long-run average) payoffs to be as good as the best expert, but not uniformly: the later the period the worse the bound. This insight is the basis of Proposition 4 below.

We first establish an upper bound for given α on how far Agent can fall short from performing as good as the best expert in the given environment.

Proposition 1 *Given discount factor δ the regret-matching strategy p_α based on past α-discounted payoffs is sequentially ε-optimal when*

$$\varepsilon = \frac{1-\alpha\delta}{1-\alpha} I \sqrt{N \frac{4\left(1-\alpha\right)^2 + (1-\delta)\alpha^2}{1-\delta\alpha^2}} + \frac{\alpha(1-\delta)}{1-\alpha} I. \tag{6}$$

All proofs are deferred to the Appendix. Looking at (6) we see that the number of experts N essentially enters with factor $\sqrt{N}$. The bound is general in the sense that it only depends on the number of experts, not on their specific strategies. Adding an expert increases the highest payoff that Agent aspires to reach, the increase is strict when she faces an environment in which this new expert is better than all the rest. An addition of any additional expert comes at the cost of strictly reducing how

close Agent can guarantee, according to (6), to be to the highest payoff among the experts. Thus, adding or removing experts may or may not be beneficial for Agent. The question of how to choose experts is not considered in this paper (see a brief discussion in Section 8).

We now show that p_α is sequentially ε-optimal for an appropriate choice of α. The value $\alpha = \alpha^*(\delta)$ is chosen to minimize $\varepsilon = \varepsilon(\alpha, \delta)$ over all $\alpha \in (0,1)$ where $\varepsilon(\alpha, \delta)$ is given in (6). To get a feeling for how α^* depends on δ when ε is small we derive approximations of the bound $\varepsilon(\alpha^*(\delta), \delta)$ when δ is close to 1. These are supplemented with approximations of $\varepsilon(\alpha, \delta)$ to highlight the trade-off between α and δ.[14]

Proposition 2 *Let $\varepsilon = \varepsilon(\alpha, \delta)$ be defined as in (6). Then*

$$\varepsilon(\alpha, \delta) = I\sqrt{N}\sqrt{2(1-\alpha) + \frac{1}{2}\frac{1-\delta}{1-\alpha}} + O\left((1-\alpha) + \frac{1-\delta}{1-\alpha}\right), \tag{7}$$

$$\varepsilon(\alpha^*(\delta), \delta) = \min_{\alpha \in (0,1)} \varepsilon(\alpha, \delta) = I\sqrt{N}\sqrt[4]{1-\delta} + 2I\sqrt{(1-\delta)} + O\left((1-\delta)^{3/4}\right), \tag{8}$$

where

$$\alpha^*(\delta) = 1 - \frac{1}{2}\sqrt{1-\delta} + O\left((1-\delta)^{3/4}\right). \tag{9}$$

In order for (6) to be small, Agent has to be very patient (δ large) and has to choose a value of decay of information $1-\alpha$ that is small in absolute terms but relatively large in comparison to $1-\delta$. Following (9), the best choice of α when δ is large is to let decay have the same magnitude as the square root of the distance between δ and 1. To gain a feeling for (8) consider δ close to 1. Note that $\frac{1}{1-\delta}$ can be interpreted as the *mean time horizon* of Agent as $(1-\delta)\sum_{t=1}^{\infty} t\delta^{t-1} = \frac{1}{1-\delta}$. Then in order to reduce the bound on maximal expected regret by 10% Agent has to increase mean time horizon by roughly 50% (as $\frac{1}{0.9^4} \approx 1.52$) and consequently increase the mean time horizon of looking into the past by roughly 25% (as $\frac{1}{0.9^2} \approx 1.23$).

We numerically calculate α^* and $\varepsilon^* = \varepsilon(\alpha^*(\delta), \delta)$ and compare these to the approximations $\hat{\alpha}$ and $\hat{\varepsilon}$ in (8) and (9) in Proposition 2 and show the values in Table 4, where we set $I = 1$.

So for instance, when there are two experts and $1 - \delta = 10^{-6}$, then we can guarantee future expected payoffs to be no worse than 0.065 as compared to those

[14]For two real-valued functions f, g we write $f = O(g)$ if there exists a constant L such that $|f(\cdot)| \leq L|g(\cdot)|$.

N	$1-\delta$	$1-\alpha^*$	$\alpha^*-\hat{\alpha}$	ε^*	$\varepsilon^*-\hat{\varepsilon}$
2	10^{-6}	5.3×10^{-4}	3×10^{-5}	0.0653	5.4×10^{-5}
2	10^{-5}	1.76×10^{-3}	1.8×10^{-4}	0.189	1.07×10^{-4}
2	10^{-4}	6×10^{-3}	0.001	0.2206	6×10^{-4}
4	10^{-6}	5×10^{-4}	0	0.0898	-1.64×10^{-3}

Table 1: Numeric examples

of the best expert. Here 0.065 can be interpreted as 6.5% of the maximal payoff difference as utility has been normalized in this table to be contained in $[0,1]$.

The literature on no-regret decision making concerns less for *expected* payoffs than providing almost sure upper bounds on the difference in payoffs. In Appendix B we present probabilistic bounds on how close Agent's discounted future payoffs are to those of the best expert. Following Cesa-Bianchi and Lugosi (2006), almost sure bounds are not available when discounting past payoffs.

5 The Role of the Rate of Adaptation

In the previous section we showed that the rate of adaptation, $1-\alpha$, has to be fine-tuned for a given discount factor δ in order to obtain Theorem 1. We now show why Theorem 1 does not hold if the rate of adaptation is too slow or too fast.

First, let us show that the rate of adaptation should be a function of δ and, as δ approaches one, $1-\alpha$ should approach zero. In other words, a strategy based on discounted past payoffs with a given rate of adaptation $1-\alpha$ independent of δ will fail to guarantee a future expected payoff arbitrarily close to that of the best expert, no matter how patient (or impatient) Agent is.

Before stating the formal result, let us show the intuition behind it. Imagine that Nature has two states, either *Rain* or *Sun*, that occur with probability 1/3 and 2/3, respectively, independently in every period. Agent receives the payoff of I if she forecasts the state of Nature correctly, otherwise she receives zero. There are two constant experts: one always forecasts *Rain*, the other always *Sun*. Given this environment, the best strategy for Agent, regardless of her discount factor, is to forecast *Sun* in each period, in other words, to always follow the recommendation

of the expert that forecasts *Sun*. This is what happens asymptotically when Agent bases her forecast on past average payoffs. Past frequencies, due to the law of large numbers, eventually reflect true probabilities and hence she will learn to forecast the more likely event. Now consider an adaptive Agent. More recent events receive more weight, and after a sufficiently long sequence of periods in which *Rain* occurred she will essentially ignore what happened before this sequence and hence forecast *Rain*. Of course, the event that such a sequence occurs has a low probability. Yet, this probability is strictly positive, thus preventing Agent from learning to forecast *Sun* in each period.

Proposition 3 *Fix $\alpha \in (0,1)$. Then there exists $\varepsilon_0 > 0$ such that for every $\delta \in (0,1)$ there does not exist a better-reply strategy based on past α-discounted payoffs that is sequentially ε_0-optimal.*

Second, let us show why it is important for the strategy to be sufficiently adaptive, in other words, what can go wrong when the rate of adaptation is too small. Consider first the canonical model in which Agent bases here future choice on *past average* payoffs. Almost all up-to-date literature (with exception of Marden et al. 2007, Mallet et al. 2009, Zapechelnyuk 2008, and Lehrer and Solan 2009) chooses this model.

More specifically, for every history h_t, the next-period mixed action of Agent is a function of $C_{1,t}$ only: $p_{t+1} = \sigma(C_{1,t})$. These strategies become decreasingly adaptive over time, their rate of adaptation is equal to $1/t$ after t periods. When some expert that has been the best so far becomes non-optimal, it may take a very long time for Agent to learn this and to start following the recommendation of a different expert. The later the period, the longer it will take Agent to adapt to changes. Thus, no matter how patient Agent is, after sufficiently many periods there will be histories such that Agent may not want to wait until her past average payoffs are able capture changes in the environment. Thus, the problem of dynamic consistency arises. After some time and some histories Agent will prefer to "forget" the past and to restart the strategy from the empty history. Therefore, these strategies fail to be dynamically consistent as defined by our concept of sequential ε-optimality.

To illustrate, let us return to our previous example and consider a non-stationary environment in which *Sun* occurs in periods 1 to m and *Rain* occurs forever thereafter. Given $T \in \mathbb{N}$, if m is sufficiently large, then Agent will forecast *Sun* in periods

$m+1, ..., m+T$ even though *Rain* occurs in each of these periods. Payoffs in periods $m+1$ to $m+T$ are equal to 0 and hence in those periods they are far from that of the best expert. So for any given discount factor δ ($\delta < 1$), one only has to choose m sufficiently large to make Agent unwilling to maintain her strategy at period $m+1$.

Proposition 4 *For every $\varepsilon < I/2$ and every $\delta \in (0,1)$ there exists $\alpha_0 < 1$ such that there does not exist a better-reply strategy based on past average or past α-discounted payoffs with $\alpha > \alpha_0$ that is sequentially ε-optimal.*

In particular, this proposition shows that none of the popular "no regret" strategies considered in the literature, referring to Hart and Mas-Colell's (2000) regret matching, $l_{\mathbf{p}}$-norm strategies of Hart and Mas-Colell (2001a) and Cesa-Bianchi and Lugosi (2003), as well as the fictitious play and its smooth variants, satisfy the objective of sequential rationality (or dynamic consistency) that is the focus of this paper.

Remark 1 Assume briefly that Agent does not discount future payoffs, but instead is concerned in each period t with average payoffs in the next T periods. Proposition 4 immediately extends. This follows directly from our example above in which we demonstrated how it can happen that Agent attains the lowest payoff in T consecutive periods when conditioning play on past average payoffs.

Similarly, our main result, Theorem 1, extends. When Agent is concerned with average payoffs in the next T periods, then the regret matching based on past α-discounted payoffs generates a sequentially ε-optimal strategy provided α is chosen appropriately and T is sufficiently large. The important underlying assumption is that the decision problem is stationary, that is, in every period Agent is concerned about the same horizon T of future payoffs.

Remark 2 We hasten to point out that if Agent faces a finitely repeated decision problem with $\overline{T}$ periods, then sequentially ε-optimal strategies fail to exist when $\varepsilon < I/2$, regardless of how past information is used. The intuition is simple. After facing $\overline{T}-1$ periods, Agent is only concerned with her payoff in the final period $\overline{T}$. Since Nature's strategy is arbitrary, the past information is irrelevant. Thus, Agent can guarantee only the maximin payoff, in our above example this is $I/2$, while the payoff of the best expert in the final round is equal to I.

6 The Role of the Discount Factor

In this paper, the discount factor is a parameter that describes the *patience* of the decision maker (who we call Agent), her intertemporal preferences that relate today's and tomorrow's utility. The statement in Theorem 1 may leave an impression that a more patient decision maker can achieve a better result in terms of discounted future payoffs. In this section we argue that this need not be true, and that the relationship between the discount factor and learning the best strategy is far more complex.

Recall that in this paper the decision maker's objective is to do as well as the best expert, and we find a more patient decision maker can get closer to the best expert. Consider now an outside observer who measures the performance of the decision maker by her long-run average payoff. What is the value for the decision maker of following the best expert from the perspective of the observer? The answer is not trivial, since an expert's discounted future payoff depends on the decision maker's discount factor, δ. When δ is higher, then maximum discounted payoff among experts can be higher when the environment is stationary, but it can be lower when the environment is non-stationary. Indeed, an expert who is best in the long run is not getting very good short-run average payoffs if the environment is changing. Therefore, it could well be that for the observer a less patient decision maker will show a better performance than a more patient one.

To illustrate, consider our example from the previous section. In every period Nature chooses *Rain* or *Sun*, the decision maker needs to forecast the state of Nature, and there are two constant experts: one always forecasts *Rain*, the other always *Sun*. Suppose that Nature deterministically alternates between m periods of *Sun* and m periods of *Rain*. To be as good as the best expert on average in the long run means here to correctly predict the state of Nature half of the time. To be as good as the best expert in the next period (i.e., when $\delta = 0$) means to correctly predict the state in each period. Of course it is impossible to perform as well as the best expert, since the strategy of Nature is unknown. It follows that an impatient decision maker aspires to a higher goal than a patient one, as she wishes to achieve a high payoff in every short run, as opposed to achieving a high average payoff in the long run. We can now explain the trade-off between focusing on long run payoffs and short run payoffs as follows. In the long run one can get arbitrarily close to the payoff of the

best expert, as her performance is based on all periods, and hence the entire past can be used to learn which expert is the best. The downside is that the long run payoff will not be very large if the environment is changing. When focusing on performance of the best expert in the short run, one has higher goals, as now one is fine-tuning the best expert to the upcoming environments, ignoring those in the distant future. The disadvantage is that it is harder to reach these goals, to get close to the best expert for the near future. The reason is that one cannot use information from the distant past as it may not be relevant. Instead one needs to focus on more recent past which essentially limits the amount of information one is gathering. This is best seen by our result that information from the recent past is not enough to learn which action is best in a stationary environment (see the example in Section 5).

Note that a higher goal may be alternatively set by adding more sophisticated experts that take into account past dependencies and adjust to changing environments. However, one has to be aware of the fact that there are many ways to condition on the past. In fact, one cannot add all experts that condition on the payoffs obtained in the previous period when infinitely many payoffs can be realized. Even when there are only finitely many payoffs, the set of all experts that condition on the past k rounds increases exponentially in k . This makes the task of selecting the set of experts particularly difficult as the precision of how close the decision maker can get to the payoff of the best expert negatively depends on the number of experts. In contrast, reducing the discount factor is a unidimensional problem that highlights in a simple way the trade-off between adapting to a changing environment and gathering sufficient information to be able to adapt.

It would be interesting to consider the framework where the decision maker sets her goals by strategically choosing the discount factor. We leave formalization and analysis of this problem for future research. Here we only note that a decision maker who is interested in long-run average payoffs may wish to decrease the discount factor away from 1, understanding the trade-off between a higher aspiration level when δ is smaller and more efficient learning when δ is larger. In applications this is done by calibrating δ to past observations, as undergone by Mallet et al. (2009).

7 Noisy Observations

In this section we return to our basic model and extend it to allow for observations of expert payoffs to be noisy. We will show that Theorem 1 continues to hold, with a slightly looser upper bound due to the additional source of error.

In our basic model, Agent observes the state of nature and computes the forgone payoff of not following the recommendation a_t^j of expert j in period t as $u(a_t^j, \omega_t)$. Suppose now that Agent does not observe states of Nature. Instead, she only observes payoffs, and these are subject to noise. Let $\tilde{u}_t(0)$ be Agent's observed payoff generated in period t and let $\tilde{u}_t(j)$ be that of expert j. We assume that

$$\begin{aligned} \tilde{u}_t(0) &= (1-\lambda)\, u_t(a_t, \omega_t) + \lambda \xi_t(0), \\ \tilde{u}_t(j) &= (1-\lambda)\, u_t(a_t^j, \omega_t) + \lambda \xi_t(j) \text{ for } j \in J. \end{aligned}$$

Here λ is a parameter that measures the *level of contamination* of Agent's information and satisfies $0 \leq \lambda < 1$. The only assumption on noisy payoffs is that they satisfy the same constraints as the true payoffs, namely for every $t = 1, 2, \ldots$ and every $j \in J \cup \{0\}$,

$$0 \leq \tilde{u}_t(j) \leq I. \tag{10}$$

In particular, no assumptions are made on the relationship of the noise of different experts in different periods.

The following examples fall within this framework.

(i) Agent's payoffs are perfectly observable, while those of experts can be noisy. So noise in observed experts' payoffs only matters for those experts who have chosen a different action. Here $\tilde{u}_t(0) = u_t(a_t, \omega_t)$ and $\tilde{u}_t(j) = u_t(a_t^j, \omega_t)$ whenever $a_t^j = a_t$. For instance, one can model the situation where with some probability bounded above by λ expert 1 acts as if he obtained the payoff of the best expert in that period (instead of her reporting his own).

(ii) Payoffs are perfectly observable, but experts possibly do not face the same state as Agent does. Here λ is the maximal probability that expert j does not face the same state as Agent in period t. The probability of facing a different state than Agent can be drawn independently for each period. It could also be that some experts simply never face the environment of Agent.

(iii) Sometimes an expert's payoffs are not observable. With probability smaller than λ the payoff of expert j is not observed in period t. In this case $\xi_t(j)$ is a part of the strategy of Agent and we set $\xi_t(j) = 0$. Analogous to (ii), there are no assumptions on how the event that the payoff of expert j is not observed in period t depends on other events. Here the most natural model is the case where the event that the payoff of expert j is observed is independent of whether payoffs of other experts were observed.

We continue to measure performance by discounting the past, the only difference is that these calculations are now based on the noisy payoffs $\tilde{u}_t(j)$. Specifically, the past α-discounted payoff of expert j is defined the same way as before, $C_{\alpha,0}(j) = 0$, and for every $t \geq 1$

$$C_{\alpha,t}(j) = \alpha C_{\alpha,t-1}(j) + (1-\alpha)\tilde{u}_t(j).$$

Future performance is still measured in terms of discounted future payoffs, here only the true utilities u_t matter. Let $U_{t,\delta}(j)$ be the discounted future payoff of expert j. The same notations with "$\sim$" refer to the corresponding expressions with noisy payoffs. Let $\xi_{t,\delta}(j) = (1-\delta)\sum_{i=0}^{\infty}\delta^i\xi_{t+i}(j)$.

There is no need for new proofs. Following Theorem 1 we know there exists ε such that $\tilde{U}_{t,\delta} \geq \tilde{U}_{t,\delta}(j) - \varepsilon$ for all t and all j. Thus, we obtain that

$$(1-\lambda)U_{t,\delta} \geq (1-\lambda)U_{t,\delta}(j) + \lambda\left(\xi_{t,\delta}(j) - \xi_{t,\delta}(0)\right) - \varepsilon$$

and hence

$$U_{t,\delta} \geq U_{t,\delta}(j) + \frac{\lambda}{1-\lambda}\left(\xi_{t,\delta}(j) - \xi_{t,\delta}(0)\right) - \frac{\varepsilon}{1-\lambda}.$$

This leads us to the next proposition, in particular we find that Theorem 1 continues to hold.

Proposition 5 *Given discount factor δ and observation of noisy payoffs, let $\theta, b_t \in [0, I]$ be such that $E\left(\xi_t(j)\,|h_{t-1}\right) \in [b_t, b_t+\theta]$ for all t and all h_{t-1} then regret-matching strategy p_α based on α-discounted past payoffs is sequentially ε-optimal when*

$$\varepsilon = \frac{1}{1-\lambda}\left(\frac{1-\alpha\delta}{1-\alpha}I\sqrt{N\frac{4(1-\alpha)^2+(1-\delta)\alpha^2}{1-\delta\alpha^2}} + \frac{\alpha(1-\delta)}{1-\alpha}\right) + \frac{\lambda\theta}{1-\lambda}. \qquad (11)$$

Note that when $\theta = 0$, the above bound is precisely the bound (6) for the noiseless environment increased by the factor of $1/(1-\lambda)$. The effect of noise is rather small:

if, for instance, the information is contaminated by 10% ($\lambda = 0.1$), then the resulting bound is greater only by about 11%. When $\theta > 0$ then there is an additional term reflecting the difference between the expected utility observed by Agent and expected utility generated for Agent.

8 Discussion and Conclusion

In conclusion, we discuss various issues related to our results.

A comment on Propositions 1 and 6. The proofs of the central result of this paper, Proposition 1, and its counterpart that concerns probabilistic bounds, Proposition 6 in Appendix B, contain new elements. As in Hart and Mas-Colell (2000), we use a quadratic potential function to bound past regret, but we cannot use the Approachability Theorems (Hart and Mas-Colell, 2001a; Lehrer, 2003) to derive our result, as they apply to averaging the past, while we discount the past. Further steps that are new in our proofs involve connecting discounted past payoffs to discounted future payoffs (in Proposition 1), and extension of the Hoeffding-Azuma inequality to infinite series and its use in deriving probabilistic bounds (Proposition 6).

Convex action sets. Our sequentially ε-optimal strategy is randomized as it specifies to adapt recommendations probabilistically whenever at least two experts have performed strictly better than Agent. However, there are applications where it is does not seem desirable to follow advice according to a lottery outcome. Instead, one would expect Agent to find a compromise in such situations. For instance, assume that Agent is a financial broker and there are two experts, *E1* and *E2*, who did equally well in the past and better than Agent herself. Suppose that for the next period expert *E1* recommends to increase holding of a certain stock by 20%, while expert *E2* recommends to do nothing. Then a reasonable action for Agent is to increase the stock holding by some amount between 0 and 20%. To make this behavior possible, two additional assumptions are necessary. First, Agent's set of actions, A, should be convex, so that "compromise" actions exist. Second, Agent should be risk averse and prefer compromises to lotteries. That is, her utility function, $u(a, \omega)$, should be (weakly) concave in a. Under these assumptions Agent choose the expected action resulting from the randomized strategy. In the above example she should increase

the stock holding by $x\%$ where x is the probability of following expert *E1* under the randomized strategy. With this modification all our results go through without changes.

Behavior of Nature independent of Agent's actions. In many situations it is not natural to assume that the state of Nature depends on previous choices of Agent. For such applications where Nature is less powerful one may wonder if our bound would look different. However this is not the case. The statements in Theorem 1 and Proposition 1 have to hold for every deterministic sequence of states of Nature, including those sequences that, by coincidence, ex post look like Nature has been conditioning the choice of its states on Agent's past actions.

Bounded-recall strategies. A different way of designing strategies that are able to adapt to changing environments is to use *bounded recall*, where Agent observes only the information from a certain number of the last periods (Zapechelnyuk, 2008; Lehrer and Solan, 2009). It is an open question whether our objective can be achieved by these strategies. Note, however, that bounded recall strategies are more computationally complex than strategies based on discounting the past. In order to implement a bounded recall strategy with length of recall m, a decision maker has to remember the information about each of the last m periods. For our strategy based on discounted past payoffs she needs to remember the accumulated discounted value of payoffs for each expert and for herself from the last period and update this with information in every period. So Agent's memory consists of $N + 1$ real numbers (in particular, she does not have to remember which period she is in).

The problem of expert selection. A special feature in this literature on learning in an unknown environment is the way in which one deals with the complexity of the environment. One cannot hope to perform well in each period, thus one compares performance to a given finite set of experts or benchmark strategies. An open question not analyzed in this paper is how to choose such experts. The more experts there are, the higher is our bound, as it increases with $\sqrt{N}$ and it does not depend on the specific types of experts. Naturally the bound can be lowered if one adds assumptions on the relationship between the experts. Note that it does not make sense to add new experts that are convex combinations of the existing experts. This is because our bound not only applies to the given finite set of experts, but also to their convex

hull. A more general analysis of the interplay between experts and the bound is not straightforward and hence is left for future research.

The multi-armed bandit setting. Consider learning under partial information where Agent observes only own payoffs but not payoffs of any other actions chosen by experts. Here we explain how to extend our algorithm to derive the same result as in Theorem 1.

Since the foregone payoffs are not observed, we use the trick of Auer et al. (1995) to construct their unbiased estimates. Define the estimate $\hat{u}_t(j)$ of a payoff of each expert j in every period t as u_t/p_t^j if expert j's recommendation is chosen in period t, and $\hat{u}_t(j) = 0$ otherwise. Then, in each period with probability $1 - \eta$ use our strategy p_α applied to the past α-discounted estimated payoffs, and with probability η follow the recommendation of a random expert, choosing each expert equally likely. These adjustments can be easily accounted for in our proofs to yield a result as in Theorem 1. The parameter $\eta > 0$ is called the rate of experimentation, its value can be fine-tuned for the best performance. Naturally, the new bound as in Proposition 1 will be larger, as now Agent conditions her decisions on much less information.

Other questions for future research. We consider learning with full information and include an extension where there are errors in observability of past payoffs. A natural extension is to consider the so-called *bandit* setting where only payoffs of the action chosen are observed (but not the forgone payoffs).[15] Another natural road for future research is to consider how our strategy performs in games. The approach in the present literature has been to get good performance in terms of learning by focusing on *conditional regrets*, which can be modeled by considering special experts that condition their play on the outcome in the past period (e.g., Hart and Mas-Colell, 2000; Hart, 2005).

Appendix A: Proofs

Below are the proofs omitted in the previous sections. In order to retain proximity to the literature we formulate proofs in terms of regrets. We use the following notation.

[15] See Hart and Mas-Colell (2001b) for similar results in the setting of *Hannan regret learning* and Foster and Young (2006) in the setting of *regret testing*.

For every period t and every $j \in J$ denote by $r_t(j, a_t, \omega_t)$ the *instantaneous regret* for not following the recommendation of expert j in that period, defined by

$$r\left(j, a_t, \omega_t\right) = u(a_t^j, \omega_t) - u(a_t, \omega_t).$$

In later proofs we will use the fact that $|r\left(j, a_t, \omega_t\right)| \leq I$ (since the utilities are in $[0, I]$). Define Agent's *discounted future regret* for expert $j \in J$ at time t_0 by

$$R_{t_0,\delta}(j) = (1-\delta)\sum\nolimits_{t=t_0}^{\infty} \delta^{t-t_0} r_t(j). \tag{12}$$

Observe that $R_{t_0,\delta}(j) = U_{t_0,\delta}(j) - U_{t_0,\delta}(0)$ for $j \in J$. Then a strategy p is sequentially ε-optimal if for every strategy q of Nature, every initial period t_0 and every initial history h_{t_0-1},

$$E\left[R_{t_0,\delta}(j)|h_{t_0-1}\right] \leq \varepsilon.$$

Let $D_{1,t}$ denote the average of past regrets of not following expert j, so

$$D_{1,t}(j) = \frac{1}{t}\sum\nolimits_{i=1}^{t} r_i(j),$$

and let $D_{\alpha,t}(j)$ be the following measure of discounted past regrets:

$$D_{\alpha,t}(j) = (1-\alpha)\sum_{i=1}^{t} \alpha^{t-i} r_i(j).$$

Observe that $D_{1,t}(j) = C_{1,t}\left(j\right) - C_{1,t}\left(0\right)$ and $D_{\alpha,t}(j) = C_{\alpha,t}\left(j\right) - C_{\alpha,t}\left(0\right)$ for all $j \in J$.

A.1 Proof of Proposition 1

Consider discounted future regrets from the perspective of time t_0. Suppose that Agent plays the regret matching strategy (5) w.r.t. discounted past regrets. Denote by $D_{\alpha,t}^{+}$ and $D_{\alpha,t}^{-}$ the positive and negative parts of $D_{\alpha,t}$, respectively, i.e., $D_{\alpha,t}^{+}(j) = \max\left\{D_{\alpha,t}(j), 0\right\}$ and $D_{\alpha,t}^{-}(j) = \min\left\{D_{\alpha,t}(j), 0\right\}$, $j \in J$. Also, denote by r_t and $D_{\alpha,t}$ the correspondent vectors of instantaneous and discounted past regrets, i.e., $r_t = (r_t(j))_{j\in J}$ and $D_{\alpha,t} = (D_{\alpha,t}(j))_{j\in J}$. The proof is divided into three steps.

Step 1. For every $t = 1, 2, \ldots$ we define $X_t = D_{\alpha,t-1}^{+} \cdot r_t$ and show that

$$E\left[X_t|h_{t-1}\right] \equiv D_{\alpha,t-1}^{+} \cdot E\left[r_t|h_{t-1}\right] = 0.$$

Step 2. Define $\rho_t^2 = \left\|D_{\alpha,t}^+\right\|^2 \equiv \left\|D_{\alpha,t} - D_{\alpha,t}^-\right\|^2$, where $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^N$. We derive an upper bound on $\left[(1-\delta)\sum_{t=t_0}^{\infty}\delta^{t-t_0}\rho_t^2\right]$ for the regret matching strategy with α-discounting. We show that

$$(1-\delta)\sum_{t=t_0}^{\infty}\delta^{t-t_0}\rho_t^2 \leq \sum_{t=t_0}^{\infty} b_t X_t + I^2 N\frac{4(1-\alpha)^2+(1-\delta)\alpha^2}{1-\delta\alpha^2}, \tag{13}$$

where $b_{t_0}, b_{t_0+1}, \ldots$ are some positive bounded coefficients.

By Step 1, for every $t \geq t_0$, $E[X_t|h_{t-1}] = 0$, and the following is immediate:

$$E\left[(1-\delta)\sum_{t=t_0}^{\infty}\delta^{t-t_0}\rho_t^2\middle| h_{t_0-1}\right] \leq I^2 N\frac{4(1-\alpha)^2+(1-\delta)\alpha^2}{1-\delta\alpha^2}. \tag{14}$$

Step 3. We show that for every $j \in J$,

$$R_{t_0,\delta}(j) \leq \frac{1-\alpha\delta}{1-\alpha}\sqrt{(1-\delta)\sum_{t=t_0}^{\infty}\delta^{t-t_0}\rho_t^2} + \frac{\alpha(1-\delta)}{1-\alpha}I.$$

Following (14) and applying Jensen's inequality, we obtain

$$\max_{j\in J} E[R_{t_0,\delta}(j)|h_{t_0-1}] \leq \frac{1-\alpha\delta}{1-\alpha}I\sqrt{N\frac{4(1-\alpha)^2+(1-\delta)\alpha^2}{1-\delta\alpha^2}} + \frac{\alpha(1-\delta)}{1-\alpha}I.$$

Proof of Step 1. Suppose that $D_{\alpha,t-1}^+(j) > 0$ for some $j \in J$ (otherwise it is immediate that $D_{\alpha,t-1}^+ \cdot r_t = 0$). In every period t the regret matching strategy p_t assigns probability $\frac{D_{\alpha,t-1}^+(j)}{\sum_{j'\in J} D_{\alpha,t-1}^+(j')}$ to action recommended by expert j. Hence, for every expert $k \in J$ and every $\omega \in \Omega$ we have

$$\begin{aligned} E[r_t(k)|h_{t-1}] &= \int_A u(a,\omega)dp_t^k(a) - \int_A u(a,\omega)dp_t(a) \\ &= \int_A u(a,\omega)dp_t^k(a) - \sum_{j\in J}\frac{D_{\alpha,t-1}^+(j)}{\sum_{j'\in J} D_{\alpha,t-1}^+(j')}\int_A u(a,\omega)dp_t^j(a). \end{aligned}$$

For short, we write $u(p_t^k,\omega)$ for $\int_A u(a,\omega)dp_t^k(a)$, $k \in J$. Therefore,

$$\begin{aligned} D_{\alpha,t-1}^+ \cdot E[r_t|h_{t-1}] &= \sum_{k\in J} D_{\alpha,t-1}^+(k)\left[u(p_t^k,\omega) - \sum_{j\in J} u(p_t^j,\omega)\frac{D_{\alpha,t-1}^+(j)}{\sum_{j'\in J} D_{\alpha,t-1}^+(j')}\right] \\ &= \sum_{k\in J} D_{\alpha,t-1}^+(k)u(p_t^k,\omega) - \sum_{j\in J} u(p_t^j,\omega)\frac{D_{\alpha,t-1}^+(j)}{\sum_{j'\in J} D_{\alpha,t-1}^+(j')}\sum_{k\in J} D_{\alpha,t-1}^+(k) \\ &= \sum_{k\in J} D_{\alpha,t-1}^+(k)u(p_t^k,\omega) - \sum_{j\in J} u(p_t^j,\omega)D_{\alpha,t-1}^+(j) = 0. \end{aligned}$$

Proof of Step 2. We have $D_{\alpha,t} = \alpha D_{\alpha,t-1} + (1-\alpha) r_t$ for every $t > 1$. Therefore,

$$\begin{aligned}\rho_t^2 &= \left\|D_{\alpha,t} - D^-_{\alpha,t}\right\|^2 \le \left\|D_{\alpha,t} - D^-_{\alpha,t-1}\right\|^2 = \left\|\alpha(D_{\alpha,t-1} - D^-_{\alpha,t-1}) + (1-\alpha)(r_t - D^-_{\alpha,t-1})\right\|^2 \\ &= \alpha^2\rho_{t-1}^2 + 2\alpha(1-\alpha)(D_{\alpha,t-1} - D^-_{\alpha,t-1})(r_t - D^-_{\alpha,t-1}) + (1-\alpha)^2(r_t - D^-_{\alpha,t-1})^2,\end{aligned}$$

where the inequality follows from $D^-_{\alpha,t}$ being the closest point to $D_{\alpha,t}$ in $\mathbb{R}^{|A|}_-$. Since the instantaneous regret in every period is bounded by I, we have $(r_t - D^-_{\alpha,t-1})^2 \le N \cdot (2I)^2 = 4I^2N$. Now, using $D_{\alpha,t-1} - D^-_{\alpha,t-1} = D^+_{\alpha,t-1}$, $D^-_{\alpha,t-1} \cdot D^+_{\alpha,t-1} = 0$, and replacing $D^+_{\alpha,t-1} r_t$ by X_t, we have

$$\rho_t^2 \le \alpha^2\rho_{t-1}^2 + 2\alpha(1-\alpha)X_t + (1-\alpha)^2 \cdot 4I^2N.$$

Repeatedly applying the above inequality to $\rho_{t-1}, \rho_{t-2}, \ldots, \rho_{t_0}$ yields

$$\rho_t^2 \le \alpha^{2(t-t_0+1)}\rho^2_{t_0-1} + 2\alpha(1-\alpha)\sum_{i=t_0}^{t}\alpha^{2(t-i)}X_i + (1-\alpha)^2 4I^2N\sum_{i=t_0}^{t}\alpha^{2(t-i)}.$$

By the fact that $\rho^2_{t_0-1} \le I^2N$, we obtain

$$\begin{aligned}\rho_t^2 &\le \alpha^{2(t-t_0+1)}I^2N + 2\alpha(1-\alpha)\sum_{i=t_0}^{t}\alpha^{2(t-i)}X_i + 4I^2N\frac{(1-\alpha)\left(1-\alpha^{2(t-t_0+1)}\right)}{1+\alpha} \\ &\le 2\alpha(1-\alpha)\sum_{i=t_0}^{t}\alpha^{2(t-i)}X_i + 4I^2N\frac{1-\alpha}{1+\alpha} + I^2N\alpha^{2(t-t_0+1)}\left(1 - 4\frac{1-\alpha}{1+\alpha}\right). \qquad (15)\end{aligned}$$

Next, consider the expression $(1-\delta)\sum_{t=t_0}^{\infty}\delta^{t-t_0}\rho_t^2$, so

$$\begin{aligned}(1-\delta)\sum_{t=t_0}^{\infty}\delta^{t-t_0}\rho_t^2 &\le \sum_{t=t_0}^{\infty} b_t X_t + 4I^2N\frac{1-\alpha}{1+\alpha}(1-\delta)\sum_{i=0}^{\infty}\delta^i \\ &\quad + I^2N\left(1 - 4\frac{1-\alpha}{1+\alpha}\right)(1-\delta)\sum_{i=0}^{\infty}\delta^i\alpha^{2(i+1)} \\ &= \sum_{t=t_0}^{\infty} b_t X_t + 4I^2N\frac{1-\alpha}{1+\alpha} + I^2N\left(1 - 4\frac{1-\alpha}{1+\alpha}\right)\frac{(1-\delta)\alpha^2}{1-\delta\alpha^2} \\ &= \sum_{t=t_0}^{\infty} b_t X_t + I^2N\frac{4(1-\alpha)^2 + (1-\delta)\alpha^2}{1-\delta\alpha^2},\end{aligned}$$

where $b_{t_0}, b_{t_0+1}, \ldots$ are the resulting coefficients on the respective X_t. The coefficient of X_i in (15) corresponding to ρ_t^2 is equal to $2\alpha(1-\alpha)\alpha^{2(t-i)}$ provided $t_0 \le i \le t$. So we obtain for every $t \ge t_0$

$$b_t = (1-\delta)\sum_{i=t}^{\infty}\delta^{t-t_0+i}2\alpha(1-\alpha)\alpha^{2i} = 2\delta^{t-t_0}(1-\delta)\frac{\alpha(1-\alpha)}{1-\delta\alpha^2}. \qquad (16)$$

Proof of Step 3. First, we have

$$\rho_t = \sqrt{\sum_{j\in J} \left(D^+_{\alpha,t}(j)\right)^2} \geq \max_{j\in J} D^+_{\alpha,t}(j) \geq \max_{j\in J} D_{\alpha,t}(j) \geq D_{\alpha,t}(j), \quad j \in J. \tag{17}$$

Next, using (17), we obtain

$$\begin{aligned}\sqrt{(1-\delta)\sum_{t=t_0}^{\infty} \delta^{t-t_0}\rho_t^2} &\geq (1-\delta)\sum_{t=t_0}^{\infty} \delta^{t-t_0}\rho_t \\ &\geq (1-\delta)\sum_{t=t_0}^{\infty} \delta^{t-t_0} D_{\alpha,t}(j), \quad j \in J. \end{aligned} \tag{18}$$

Recall that $R_{t_0,\delta} = (1-\delta)(r_{t_0} + \delta r_{t_0+1} + \delta^2 r_{t_0+2} + \ldots)$. Thus, by rearranging the summands, we obtain

$$\begin{aligned}(1-\delta)\sum_{t=t_0}^{\infty} \delta^{t-t_0} D_{\alpha,t} &= (1-\delta)((1-\alpha)r_{t_0} + \alpha D_{\alpha,t_0-1}) \\ &\quad + (1-\delta)\delta((1-\alpha)(r_{t_0+1} + \alpha r_{t_0}) + \alpha^2 D_{\alpha,t_0-1}) \\ &\quad + (1-\delta)\delta^2((1-\alpha)(r_{t_0+2} + \alpha r_{t_0+1} + \alpha^2 r_{t_0}) + \alpha^3 D_{\alpha,t_0-1}) \\ &\quad + \ldots \\ &= (1-\alpha)(1-\delta)(r_{t_0} + \delta r_{t_0+1} + \delta^2 r_{t_0+2} + \ldots) \\ &\quad + (1-\alpha)\alpha\delta(1-\delta)(r_{t_0} + \delta r_{t_0+1} + \delta^2 r_{t_0+2} + \ldots) \\ &\quad + \ldots + (1-\delta)\alpha(D_{\alpha,t_0-1} + \alpha\delta D_{\alpha,t_0-1} + (\alpha\delta)^2 D_{\alpha,t_0-1} + \ldots) \\ &= ((1-\alpha)R_{t_0,\delta} + \alpha(1-\delta)D_{\alpha,t_0-1})\sum_{i=0}^{\infty}(\alpha\delta)^i.\end{aligned}$$

Next, using $\sum_{i=0}^{\infty}(\alpha\delta)^i = 1/(1-\alpha\delta)$ yields

$$\begin{aligned}(1-\delta)\sum_{t=t_0}^{\infty} \delta^{t-t_0} D_{\alpha,t} &= \frac{1-\alpha}{1-\alpha\delta} R_{t_0,\delta} + \frac{\alpha(1-\delta)}{1-\alpha\delta} D_{\alpha,t_0-1} \\ &\geq \frac{1-\alpha}{1-\alpha\delta} R_{t_0,\delta} - \frac{\alpha(1-\delta)}{1-\alpha\delta} I. \end{aligned} \tag{19}$$

Step 3 is immediate by (18) and (19). **End of Proof**.

A.2 Proof of Proposition 2

The proof is straightforward but tedious so we here only show how to verify the claims. Set $y = 1-\delta$, $x = 1-\alpha$ and $z = y/x$.

We show how to prove (7). Let $g(x,z)$ be the difference between the expression given in (6) and the first term in (7). So we wish to show that $g(x,z) = O(x+z)$ which is established by verifying the following: $g(0,0) = 0$, $\frac{d}{dx}g(x,z)$ is bounded for each given z and $\frac{d}{dz}g(x,z)$ is bounded for each given x.

We now show how to derive (9). Replace in (7) each appearance of the symbol O with a different constant and then take the derivative with respect to x. Then show that these first order conditions have a root when $x = \left(\frac{1}{2} + w\right)\sqrt{y}$ for $w = O\left(\sqrt[4]{y}\right)$ which yields $x = \frac{1}{2}\sqrt{y} + O\left(y^{3/4}\right)$.

The asymptotic bound given in (8) is derived by taking the difference between the expression in (6) and $\sqrt{2N}\sqrt[4]{1-\delta} + 2\sqrt{1-\delta}$, setting $x = \frac{1}{2}\sqrt{y}$ and then expanding with respect to y.

A.3 Proof of Proposition 3

The proposition is proven by example. Normalize utilities such that $I = 1$. Consider two actions H and T, two states H and T and payoffs given by $u(a,a') = 1$ if $a = a'$ and $u(a,a') = 0$ if $a \neq a'$. There are two experts, labeled H and T, who forecast constant actions, H and T, respectively. Suppose that states H and T are realized with probability π and $1-\pi$, respectively, independently in all periods.

Fix $\alpha < 1$ and consider any better-reply strategy p of Agent based on α-discounted past regret. Note that $r(T, a_t, \omega_t) = 1$ if $a_t = H$ and $\omega_t = T$ and $r(T, a_t, \omega_t) = -1$ if $a_t = \omega_t = H$ and $r(T, a_t, \omega_t) = 0$ otherwise. So the regret for not choosing T only depends on states that realize in rounds in which H is chosen. Recall that $D_{\alpha,t}(T) = (1-\alpha)\sum_{i=0}^{t}\alpha^{t-i}r(T, a_t, \omega_t)$. We derive a lower bound on $\Pr\left(D_{\alpha,t}(T) > 0 | n_t(H) = n\right)$ where $n_t(H) = \#\left\{t' \leq t : a_{t'} = H\right\}$.

Fix $m \leq n_t(H)$. Let $\chi_1, \chi_2, \ldots, \chi_{n_t(H)}$ be the subsequence of periods up to t where Agent played H. The probability that regrets are equal to 1 in the m most recent periods in which H was played is equal to $(1-\pi)^m$. If regrets are equal to 1 in the m most recent periods, then $D_{\alpha,t}(T)$ is smallest if all previous regrets are equal to -1. For $t \geq m$ we verify that

$$(1-\alpha)\sum_{i=1}^{n_t(H)-m}(-1)\alpha^{t-\chi_i} + (1-\alpha)\sum_{i=n_t(H)-m+1}^{n_t(H)}\alpha^{t-\chi_i} \geq \alpha^{n_t(H)}\left(1-2\alpha^m\right).$$

So if we choose $m = m(\alpha)$ such that $1 - 2\alpha^{m(\alpha)} > 0$ and if $n_t(H) \geq m(\alpha)$ then

$$\Pr\left(D_{\alpha,t}(T) > 0 | n_t(H) \geq m(\alpha)\right) \geq (1-\pi)^{m(\alpha)}.$$

In fact, our above calculations show that

$$\Pr\left(D_{\alpha,t}(T) > 0 | n_t(H) > 0\right) \geq (1-\pi)^{m(\alpha)}.$$

Similarly we verify that $\Pr\left(D_{\alpha,t}(H) < 0 | n_t(T) > 0\right) \geq \pi^{m(\alpha)}$. Moreover, conditional on t and on $n_t(H)$, $D_{\alpha,t}(T)$ and $D_{\alpha,t}(H)$ are independent random variables. Hence

$$\Pr\left(D_{\alpha,t}(T) > 0 > D_{\alpha,t}(H) | 0 < n_t(H) < t\right) \geq \left(\pi(1-\pi)\right)^{m(\alpha)}.$$

Consider a path on which agent plays both H and T. If $D_{\alpha,t}(T) > 0 > D_{\alpha,t}(H)$ then by the better reply property $a_{t+1} = T$. We have thus put a lower bound on the probability of choosing action T where this lower bound does not depend on δ. Assume that $\pi > 1/2$. This means that H is the better action, then

$$R_{1,\delta}(T) \geq \left(\pi(1-\pi)\right)^{m(\alpha)}(2\pi - 1).$$

which is a strictly positive lower bound that does not depend on δ.

In order to get around the final case in which agent plays H in all rounds we assume that nature chooses before period 1 equally likely $\pi \in \{0.4, 0.6\}$. All bounds above are cut in half which does not change the result to be proven.

A.4 Proof of Proposition 4

The proposition is proven by example. Normalize utilities such that $I = 1$. Consider the example used in the proof of Proposition 3. Assume w.l.o.g. that Agent chooses H in period 1 with probability at least $1/2$. Fix an integer m and let Nature select state H in periods $t = 1, \ldots, m$. If Agent have chosen H in period 1, then $r_1(H) = 0$ and $r_1(T) = -1$. By the better reply condition, Agent will choose H in period 2 and analogously also in all periods $3 \leq t \leq m+1$.

Consider first the past-average payoff criterion. Note that $mD_{1,m}(T) = -m$ and $mD_{1,m}(H) = 0$. In periods $t = m+1, m+2, \ldots, 2m$ let Nature choose state T. Then $(m+1)D_{1,m+1}(T) = -m+1$ and Agent continues to choose H up to period $2m$, and only in period $2m$ her past average regret for T becomes zero, $D_{1,2m}(T) = 0$.

Let us now evaluate the discounted future regret at period $m+1$. Since $r_t(T) = 1$ for $m+1 \leq t \leq 2m$, we obtain

$$R_{m+1,\delta}(T) = (1-\delta)\sum_{t=m+1}^{2m} \delta^{t-m-1} + \delta^m R_{2m+1,\delta}(T) \geq 1 - 2\delta^m.$$

Hence, given $\delta < 1$ and $\varepsilon > 0$, if m is sufficiently large, then $R_{m+1,\delta}(T) > 1 - \varepsilon$.

Now consider the past α-discounted payoff criterion. It can be verified that in this case Agent will choose H in periods $m+1, m+2, \ldots, 2m-1$ if $\alpha = \alpha(m)$ is close enough to 1. Hence, $R_{m+1,\delta}(T) \geq 1 - 2\delta^{m-1} > \varepsilon$ if m is sufficiently large, which completes the proof.

Appendix B: Probabilistic Bounds

The literature on regret-minimizing decision making is concerned with almost sure upper bounds on maximum regret that Agent may accumulate during the play. As noted by Cesa-Bianchi and Lugosi (2006), these bounds will never be small when the strategy is based on α-discounted past payoffs. This is because the overweighing of the last observation adds to the process a positive variance that never vanishes.

The goal of this section is to provide upper bounds on Agent's realized past α-discounted payoff, as well as on realized future δ-discounted payoff with a given probability (or confidence level) $\gamma < 1$. Define

$$\varepsilon^P(\alpha;\gamma) = I\sqrt{N}\sqrt{2\sqrt{-\frac{1-\alpha}{1+\alpha}\ln(1-\gamma)} + 4\frac{1-\alpha}{1+\alpha} + \alpha^{2(t-t_0)}}.$$

and

$$\varepsilon^F(\alpha,\delta;\gamma) = I\sqrt{N}\frac{1-\alpha\delta}{1-\alpha}\sqrt{\frac{2\alpha(1-\alpha)}{1-\delta\alpha^2}\sqrt{-2\frac{1-\delta}{1+\delta}\ln(1-\gamma)} + \frac{4(1-\alpha)^2 + (1-\delta)\alpha^2}{1-\delta\alpha^2}} + \frac{\alpha(1-\delta)}{1-\alpha}I.$$

Proposition 6 *Suppose that Agent has discount factor δ and uses the regret-matching strategy based on past α-discounted payoffs. Then for every time t, every history h_{t-1} and every strategy q of Nature,*

$$\Pr\left[C_{\alpha,t}(0) \geq \max_{j=1,\ldots,N} C_{\alpha,t}(j) - \varepsilon^P(\alpha;\gamma)\right] \geq \gamma \tag{20}$$

and

$$\Pr\left[U_{t,\delta}(p_\alpha) \geq \max_{j=1,\ldots,N} U_{t,\delta}(p^j) - \varepsilon^F(\alpha,\delta;\gamma)\right] \geq \gamma. \tag{21}$$

We wish to consider these bounds when δ is close to 1. Assume that α is chosen to minimize the bound on expected payoff given in (6), so $\alpha^* = 1 - \frac{1}{2}\sqrt{1-\delta} + O\left((1-\delta)^{3/4}\right)$. Then

$$\varepsilon^F = I\sqrt{N}\sqrt{2+\sqrt{-\ln(1-\gamma)}}\sqrt[4]{1-\delta} + 2I\sqrt{1-\delta} + O\left((1-\delta)^{\frac{3}{4}}\right)$$

and for t_0 large enough

$$\varepsilon^P = I\sqrt{N}\sqrt[4]{-\ln(1-\gamma)}\sqrt[8]{1-\delta} + O\left((1-\delta)^{\frac{3}{8}}\right).$$

These bounds are easily verified.

To proof Proposition 6 we first extend the Hoeffding-Azuma inequality (Hoeffding, 1963; Azuma, 1967) to infinite sums of dependent bounded random variables centered at conditional expectation.

Lemma 1 *Let $Z_1, Z_2, \ldots$ be an infinite sequence of random variables that satisfy $a_t \leq Z_t \leq b_t$ for every t. Then for every $\varepsilon > 0$,*

$$\Pr\left[\sum_{t=1}^{\infty}(Z_t - E[Z_t|Z_{t-1},\ldots,Z_1]) \geq \varepsilon\right] \leq \exp\left(-\frac{2\varepsilon^2}{\sum_{t=1}^{\infty}(b_t-a_t)^2}\right). \tag{22}$$

Proof of Lemma 1. Suppose that $\sum_{t=1}^{\infty}(b_t-a_t)^2 < \infty$ (otherwise inequality (22) holds trivially). It is sufficient to prove the claim when $a_t \leq 0 \leq b_t$ and $E(Z_t|Z_{t-1},...,Z_1) = 0$ holds for all t. Define $Z'_t = Z_t - E[Z_t|Z_{t-1},\ldots,Z_1]$. Following Hoeffding (1963, Theorem 2 and p.18), we obtain for every T_0 that

$$\Pr\left[\sum_{t=1}^{T_0} Z'_t \geq \varepsilon\right] \leq \exp\left(-\frac{2\varepsilon^2}{\sum_{t=1}^{T_0}(b_t-a_t)^2}\right). \tag{23}$$

We now extend (23) to $T = \infty$. Using Chebyscheff's inequality we obtain

$$\Pr\left[\sum_{t=T_0}^{\infty} Z'_t \geq \varepsilon\right] \leq \sum_{t=T_0}^{\infty}\Pr\left[Z'_t \geq \varepsilon\right] \leq \frac{1}{\varepsilon^2}\sum_{t=T_0}^{\infty}(b_t-a_t)^2.$$

Therefore,

$$\begin{aligned}\Pr\left[\sum_{t=1}^{\infty} Z'_t \geq \varepsilon\right] &\leq \Pr\left[\sum_{t=1}^{T_0} Z'_t \geq \varepsilon\right] + \Pr\left[\sum_{t=T_0+1}^{\infty} Z'_t \geq \varepsilon\right] \\ &\leq \exp\left(-\frac{2\varepsilon^2}{\sum_{t=1}^{T_0}(b_t-a_t)^2}\right) + \frac{1}{\varepsilon^2}\sum_{t=T_0+1}^{\infty}(b_t-a_t)^2. \end{aligned} \tag{24}$$

Taking the limit as T_0 tends to infinity, using the assumption that $\sum_{t=1}^{\infty}(b_t-a_t)^2<\infty$, and hence $\lim_{T_0\to\infty}\sum_{t=T_0}^{\infty}(b_t-a_t)^2=0$, we obtain (22). **End of Proof.**

Proof of Proposition 6. By (17) we have

$$\max_{j\in J} D_{\alpha,t}(j)\leq\sqrt{\rho_t^2} \tag{25}$$

and by Step 3 of the proof of Proposition 1 we have

$$\max_{j\in J} R_{t_0,\delta}(j)\leq\frac{1-\alpha\delta}{1-\alpha}\sqrt{(1-\delta)\sum_{t=t_0}^{\infty}\delta^{t-t_0}\rho_t^2}+\frac{\alpha(1-\delta)}{1-\alpha}I. \tag{26}$$

To obtain the bounds in (20) and (21), we will bound the probability of events $\left\{\sqrt{\rho_t^2}\geq\varepsilon\right\}$ and $\left\{(1-\delta)\sum_{t=t_0}^{\infty}\delta^{t-t_0}\rho_t^2\geq\varepsilon\right\}$ using the extended Hoeffding-Azuma inequality from Lemma 1.

First, let us deal with $\sqrt{\rho_t^2}$. By (15) we have

$$\rho_t^2\leq\sum_{i=0}^{t-t_0}b_i'X_{t-i}+4I^2N\frac{1-\alpha}{1+\alpha}+I^2N\alpha^{2(t-t_0+1)}\left(1-4\frac{1-\alpha}{1+\alpha}\right), \tag{27}$$

where

$$b_i'=2\alpha(1-\alpha)\alpha^{2i},\quad i=0,1,\ldots,t-t_0.$$

Since $|b_i'X_{t-i}|\leq b_i'I^2N$ and

$$\begin{aligned}\sum_{i=0}^{t-t_0}(b_i')^2&=4\alpha^2(1-\alpha)^2\sum_{i=0}^{t-t_0}\alpha^{4i}=4\alpha^2(1-\alpha)^2\frac{1-\alpha^{4(t-t_0+1)}}{1-\alpha^4}\\&\leq\frac{4\alpha^2(1-\alpha)}{(1+\alpha)(1+\alpha^2)}\leq\frac{2(1-\alpha)}{1+\alpha},\end{aligned}$$

by Lemma 1 we obtain

$$\Pr\left[\sum_{i=0}^{t-t_0}b_i'X_{t-i}\geq\varepsilon\right]\leq\exp\left(\frac{2\varepsilon^2}{\sum_{i=0}^{t-t_0}(2b_i'I^2N)^2}\right)\leq\exp\left(\frac{\varepsilon^2(1+\alpha)}{2I^4N^2(1-\alpha)}\right).$$

Next, from (27)

$$\begin{aligned}\Pr\left[\sum_{i=0}^{t-t_0}b_i'X_{t-i}\geq\varepsilon\right]&\geq\Pr\left[\rho_t^2\geq\varepsilon+4I^2N\frac{1-\alpha}{1+\alpha}+I^2N\alpha^{2(t-t_0+1)}\left(1-4\frac{1-\alpha}{1+\alpha}\right)\right]\\&\geq\Pr\left[\rho_t^2\geq\varepsilon+4I^2N\frac{1-\alpha}{1+\alpha}+I^2N\alpha^{2(t-t_0+1)}\right]\end{aligned}$$

Let $1-\gamma = \exp\left(\frac{\varepsilon^2(1+\alpha)}{2I^4N^2(1-\alpha)}\right)$, then

$$\varepsilon = \sqrt{-2I^4N^2\frac{1-\alpha}{1+\alpha}\ln(1-\gamma)} = I^2N\sqrt{-2\frac{1-\alpha}{1+\alpha}\ln(1-\gamma)}.$$

Hence, we obtain

$$\Pr\left[\rho_t^2 \geq I^2N\sqrt{-2\frac{1-\alpha}{1+\alpha}\ln(1-\gamma)} + 4I^2N\frac{1-\alpha}{1+\alpha} + I^2N\alpha^{2(t-t_0+1)}\right] \leq 1-\gamma,$$

and inequality (20) is straightforward by (25).

Now, let us deal with $(1-\delta)\sum_{t=t_0}^{\infty}\delta^{t-t_0}\rho_t^2$. By Step 2 of the proof of Proposition 1, we have

$$(1-\delta)\sum_{t=t_0}^{\infty}\delta^{t-t_0}\rho_t^2 \leq \sum_{t=t_0}^{\infty} b_t X_t + I^2N\frac{4(1-\alpha)^2+(1-\delta)\alpha^2}{1-\delta\alpha^2}, \tag{28}$$

where from (16) coefficients b_t satisfy

$$b_t = 2\delta^{t-t_0}(1-\delta)\frac{\alpha(1-\alpha)}{1-\delta\alpha^2}.$$

Since $|b_tX_t| \leq b_tI^2N$ and

$$\begin{aligned}\sum_{t=t_0}^{\infty} b_t^2 &= \left(2(1-\delta)\frac{\alpha(1-\alpha)}{1-\delta\alpha^2}\right)^2\sum_{t=t_0}^{\infty}\delta^{2(t-t_0)} = \left(2(1-\delta)\frac{\alpha(1-\alpha)}{1-\delta\alpha^2}\right)^2\frac{1}{1-\delta^2}\\ &= \left(2\frac{\alpha(1-\alpha)}{1-\delta\alpha^2}\right)^2\frac{1-\delta}{1+\delta},\end{aligned}$$

by Lemma 1 we obtain

$$\Pr\left[\sum_{t=t_0}^{\infty} b_tX_t \geq \varepsilon\right] \leq \exp\left(\frac{2\varepsilon^2}{\sum_{i=t_0}^{\infty}(2b_tI^2N)^2}\right) = \exp\left(2\varepsilon^2\frac{1+\delta}{I^4N^2(1-\delta)}\left(\frac{1-\delta\alpha^2}{4\alpha(1-\alpha)}\right)^2\right).$$

Next, from (28)

$$\Pr\left[\sum_{t=t_0}^{\infty} b_tX_t \geq \varepsilon\right] \geq \Pr\left[(1-\delta)\sum_{t=t_0}^{\infty}\delta^{t-t_0}\rho_t^2 \geq \varepsilon + I^2N\frac{4(1-\alpha)^2+(1-\delta)\alpha^2}{1-\delta\alpha^2}\right].$$

Let $1-\gamma = \exp\left(2\varepsilon^2\frac{1+\delta}{I^4N^2(1-\delta)}\left(\frac{1-\delta\alpha^2}{4\alpha(1-\alpha)}\right)^2\right)$, then

$$\varepsilon = I^2N\frac{2\alpha(1-\alpha)}{1-\delta\alpha^2}\sqrt{-2\frac{1-\delta}{1+\delta}\ln(1-\gamma)}.$$

Hence, we obtain

$$\Pr\left[(1-\delta)\sum_{t=t_0}^{\infty}\delta^{t-t_0}\rho_t^2 \geq I^2N\frac{2\alpha(1-\alpha)}{1-\delta\alpha^2}\sqrt{-2\frac{1-\delta}{1+\delta}\ln(1-\gamma)} + I^2N\frac{4(1-\alpha)^2+(1-\delta)\alpha^2}{1-\delta\alpha^2}\right] \leq 1-\gamma,$$

and inequality (21) is straightforward by (26). **End of Proof**

References

Auer, P., N. Cesa-Bianchi, Y. Freund, and R. E. Schapire (1995). Gambling in a rigged casino: the adversarial multi-armed bandit problem. In *Proceedings of the 36th Annual Symposium on Foundations of Computer Science*, pp. 322–331.

Auer, P. and P. M. Long (1999). Structural results about on-line learning models with and without queries. *Machine Learning 36*, 147–181.

Azuma, K. (1967). Weighted sums of certain dependent random variables. *Tôhoku Mathematical Journal 19*, 357–367.

Brown, G. (1951). Iterative solutions of games by fictitious play. In T. Koopmans (Ed.), *Activity Analysis of Production and Allocation*, Volume 13 of *Cowles Commission Monograph*, pp. 374–376. New York: Wiley.

Cesa-Bianchi, N., Y. Freund, D. P. Helmbold, and M. K. Warmuth (1996). On-line prediction and conversion strategies. *Machine Learning 25*, 71–110.

Cesa-Bianchi, N. and G. Lugosi (2003). Potential-based algorithms in on-line prediction and game theory. *Machine Learning 51*, 239–261.

Cesa-Bianchi, N. and G. Lugosi (2006). *Prediction, Learning, and Games*. Cambridge University Press.

Cesa-Bianchi, N., Y. Mansour, and G. Stoltz (2007). Improved second-order bounds for prediction with expert advice. *Machine Learning 66*, 321–352.

Erev, I. and A. E. Roth (1998). Prediction how people play games: Reinforcement learning in games with unique strategy equilibrium. *American Economic Review 88*, 848–881.

Foster, D. and R. Vohra (1993). A randomization rule for selecting forecasts. *Operations Research 41*, 704–709.

Foster, D. and R. Vohra (1997). Calibrated learning and correlated equilibrium. *Games and Economic Behavior 21*, 40–55.

Foster, D. and R. Vohra (1998). Asymptotic calibration. *Biometrika 85*, 379–390.

Foster, D. and R. Vohra (1999). Regret in the online decision problem. *Games and Economic Behavior 29*, 7–35.

Foster, D. and H. P. Young (2006). Regret testing: learning to play Nash equilibrium without knowing you have an opponent. *Theoretical Economics 1*, 341–367.

Freund, Y. and R. Schapire (1999). Adaptive game playing using multiplicative weights. *Games and Economic Behavior 29*, 79–103.

Fudenberg, D. and D. Levine (1995). Universal consistency and cautious fictitious play. *Journal of Economic Dynamics and Control 19*, 1065–1089.

Fudenberg, D. and D. Levine (1999). Conditional universal consistency. *Games and Economic Behavior 29*, 104–130.

Gordon, G. J., A. Greenwald, and C. Marks (2008). No-regret learning in convex games. Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland.

Greenwald, A. and A. Jafari (2003). A general class of no-regret learning algorithms and game-theoretic equilibria. In *Proceedings of the 2003 Computational Learning Theory Conference*, pp. 1–11.

Hannan, J. (1957). Approximation to Bayes risk in repeated play. In M. Dresher, A. W. Tucker, and P. Wolfe (Eds.), *Contributions to the Theory of Games, Vol. III*, Annals of Mathematics Studies 39, pp. 97–139. Princeton University Press.

Hart, S. (2005). Adaptive heuristics. *Econometrica 73*, 1401–1430.

Hart, S. and A. Mas-Colell (2000). A simple adaptive procedure leading to correlated equilibrium. *Econometrica 68*, 1127–1150.

Hart, S. and A. Mas-Colell (2001a). A general class of adaptive procedures. *Journal of Economic Theory 98*, 26–54.

Hart, S. and A. Mas-Colell (2001b). A reinforcement procedure leading to correlated equilibrium. In G. Debreu, W. Neuefeind, and W. Trockel (Eds.), *Economic Essays: A Festschrift for Werner Hildenbrand*, pp. 181–200. Springer.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of American Statistical Association 58*, 13–30.

Lehrer, E. (2003). A wide range no-regret theorem. *Games and Economic Behavior 42*, 101–115.

Lehrer, E. and E. Solan (2009). Approachability with bounded memory. *Games and Economic Behavior*. Forthcoming.

Littlestone, N. and M. Warmuth (1994). The weighted majority algorithm. *Information and Computation 108*, 212–261.

Mailath, G. J., A. Postlewaite, and L. Samuelson (2005). Contemporaneous perfect epsilon-equilibria. *Games and Economic Behavior 53*, 126–140.

Mallet, V., G. Stoltz, and B. Mauricette (2009). Ozone ensemble forecast with machine learning algorithms. *Journal of Geophysical Research 114*.

Marden, J. R., G. Arslan, and J. S. Shamma (2007). Regret based dynamics: convergence in weakly acyclic games. In *Proceedings of the Sixth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS07)*, Honolulu, Hawaii, USA, pp. 194–201.

Radner, R. (1980). Collusive behaviour in noncooperative epsilon-equilibria of oligopolies with long but finite lives. *Journal of Economic Theory 22*, 136–154.

Ray, D. and R. Wang (2001). On some implications of backward discounting. New York University, mimeo.

Roth, A. E. and I. Erev (1995). Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. *Games and Economic Behavior 8*, 164–212.

Vovk, V. (1998). A game of prediction with expert advice. *Journal of Computer and System Sciences 56*, 153–173.

Zapechelnyuk, A. (2008). Better-reply dynamics with bounded recall. *Mathematics of Operations Research 33*, 869–879.